Purchase Signatures of Retail Customers

Clément Gautrais, René Quiniou, Peggy Cellier, Thomas Guyet, Alexandre Termier

Purchase Signatures of Retail Customers

Clement Gautrais[1], René Quiniou[2], Peggy Cellier[3], Thomas Guyet[4], and Alexandre Termier[1]

[1] University of Rennes 1, IRISA, France
[2] Inria Rennes, IRISA, France
[3] INSA Rennes, IRISA, France
[4] Agrocampus Ouest, IRISA, France

Abstract. In the retail context, there is an increasing need for understanding individual customer behavior in order to personalize marketing actions. We propose the novel concept of *customer signature*, that identifies a set of important products that the customer refills regularly. Both the set of products and the refilling time periods give new insights on the customer behavior. Our approach is inspired by methods from the domain of sequence segmentation, thus benefiting from efficient exact and approximate algorithms. Experiments on a real massive retail dataset show the interest of the signatures for understanding individual customers.

1 Introduction

Retail, and more specifically understanding the behavior of supermarket customers, has been a strong motivation for data mining researchers since the early 1990s. Several methods have been developed in this field, such as mining frequent itemset [1], frequent sequential patterns [2] or more recently high utility itemsets [3]. These methods discover sets of products that are bought together in a large enough number of tickets, possibly with some extra information (e.g. sequencing, utility). They can be exploited to understand (large) *groups of customers*. However, with the success of loyalty programs and the increasing number of customers shopping at online grocery pick-up, a promising trend is "personalized marketing". This requires a fine grained understanding of the purchasing behavior of individual customers, in order to make relevant personalized suggestions.

In this context of personalized marketing, an important information is the "rhythm" of the individual customer. The main idea is to identify the set of products that the customer always wants to have stocked at home, and that she will thus buy on a more or less regular basis. The rhythm corresponds to the "refilling period". Extracting such information may help analysts to get insights about their customers in order to design personalized marketing campaigns. In practice, the problem is to discover from the set of the customer receipts, a set of products that are regularly purchased. A difficulty is that all the products that the customer wants to have in stock are not likely to be bought at the same time:

depending on depletion rates, renewing all such products will be distributed over several receipts.

Recently, Customer Relationship Management (CRM) has manifested interest in data mining [4], but with a focus on clustering techniques, for example to characterize segment profiles [5]. However, clustering cannot uncover the flexible time regularity of customers' purchases since time periods have to be fixed in advance. Most existing itemset mining algorithms [1, 6] consider only sets of products that are bought on a single receipt, and cannot be used for this problem. Periodic patterns [7] can find regularities through a sequence of receipts however they extract patterns with strict temporal period. Mannila et al. [8] have proposed parallel episodes to extract temporal regularities but the approach requires fixed predefined equal size windows over the sequence of events, which lacks flexibility for the problem at hand. In [9], Casas-Garriga defined a method that also adapts the window size to the data, however it still requires a maximal time interval between two events.

In this paper, we define the problem of extracting *customer signature* through a sequence of receipts. A customer signature represents a maximal set of products that are bought regularly, possibly in several receipts and such that the regularity is not strict. We show that this problem can be formalized as a sequence segmentation problem. There is an important literature about sequence segmentation. We have adapted the formal setting provided in Bingham's survey [10]. The most significant adaptation lies in the notion of segment representatives that represent occurrences of a common set of products and on the distance from sequence elements to their related representatives. Roughly, we shift from a local error view to a global one (see Section 2 for more details).

The contributions of this paper are threefold. First, the signature mining problem is defined as a segmentation problem allowing to take advantage of the many algorithms that have been proposed in the sequence segmentation field (Section 3.1). Second, we have adapted and evaluated an algorithm based on dynamic programming for sequence segmentation [11] which gives an exact solution (Section 3.2). Third, a thorough experimental study on real massive supermarket data shows the interest of our approach (Section 4).

2 Background

This section provides the data mining vocabulary used in the sequel, and presents briefly the well-studied sequence segmentation problem.

In pattern mining, an *itemset* T is a set of literals called *items*. Let $\mathcal{I}$ be the set of all items. A *sequence* α is an ordered list of itemsets, denoted by $\alpha = \langle T_1, \ldots, T_m \rangle$. In the retail context, a receipt is an itemset (a set of purchased products) and a *customer purchase sequence* is a sequence of receipts identifying the products bought by a customer at each of her visits to a supermarket during the analysis period. For instance, $\langle (p_1,\ p_2)\ (p_3)\ (p_1)\ (p_4,\ p_2,\ p_3)\ (p_1) \rangle$ is a sequence of five receipts where four different products are bought one or several times. A receipt may have an associated timestamp which indicates the purchase

date. We assume that the timestamp of T_k is implicitly the index of T_k, *i.e.* k. By extension, the timestamp of any product of a receipt T_k is the timestamp associated with this receipt T_k. A *customer sequence database SDB* is a set of tuples (C_{id}, α) where C_{id} is a customer identifier and α is the sequence of her receipts.

Our proposal is grounded on sequence or time series segmentation which has received much attention in the literature. In [10], the segmentation problem is formulated as follows. Let $\alpha = \langle T_1, T_2, ..., T_n \rangle$ be a d-dimensional sequence where $T_i \in \mathbb{R}^d$. A k-segmentation S of α is a partition of α into k non-overlapping contiguous subsequences called *segments*, i.e. $S = \langle S_1, S_2, ..., S_k \rangle$ and $\forall i \in 1 \dots k$, $S_i = \langle T_{b(i)}, \dots, T_{b(i+1)-1} \rangle$, where $b(i)$ is the index of the first element of the i-th segment. A segmentation associates a *representative*, $\mu(S_i)$, with each segment by aggregating the values of the segment. Generally $\mu(S_i)$ is a single value such as mean or median, or a pair of values such as (min, max) or (mean, slope). This reduction results in a loss of information in the sequence representation which can be measured by the reconstruction error defined as:

$$E_p(\alpha, S) = \sum_{S_i \in S} \sum_{T \in \alpha} dist(T, \mu(S_i))^p$$

where $dist(T, \mu(s))$ represents the distance between the d-dimensional point T and the representative of the segment it belongs to. The p parameter refers to the Lp norm. In practice, the median ($p = 1$) or the mean ($p = 2$) usually serves as segment representatives. The *segmentation problem* consists in finding the segmentation that minimizes the reconstruction error:

$$S_{opt}(\alpha, k) = \underset{S \in \mathscr{S}_{n,k}}{\arg\min}\, E_p(\alpha, S)$$

where $\mathscr{S}_{n,k}$ represents the set of all k-segmentations of sequences of length n.

3 Mining Signatures

In this section, we present the signature mining problem in the sequence segmentation framework. Indeed, mining a signature from a customer purchase sequence α can be seen as segmenting α into k non-overlapping and non-empty segments that cover all receipts from α and such that every segment contains a common maximal subset of products, called the *customer signature*. The signature mining problem can thus be fitted to the segmentation problem providing the opportunity to use the many exact or approximate algorithms that have been proposed in the sequence segmentation field.

3.1 Mining Signatures with Sequence Segmentation

Section 2 introduced the problem of segmenting a sequence $\alpha = \langle T_1, T_2, ..., T_n \rangle$ into k segments. Let $\mathscr{S}_{n,k}$ denote the set of all k-segmentations of a sequence

α of length n and $S = \langle S_1, S_2, ..., S_k \rangle$ be an element of $\mathscr{S}_{n,k}$. Following the representation proposed in SPAM [12], a receipt can be represented by a bitmap[5] of dimension d such that if item i_j belongs to the receipt then the j-th bit of the bitmap is set to 1, otherwise the j-th bit is set to 0. The representative r_i of a segment is then defined as the set of items that belongs to at least one receipt in the segment, i.e. the union of the segment receipts. This union can be computed by a boolean disjunction on bitmaps: $r_i = \bigvee_{t \in S_i} t$. The k-signature of a purchase sequence is the set of items that are common to every segments from a segmentation of size k, so it corresponds to the intersection of the k segment representatives. It can be computed by a boolean intersection of the related bitmaps: $Sig_k(\alpha, S) = \bigwedge_{j=1}^{k} r_j$. As we intend to represent a customer purchase sequence by its signature, the reconstruction error is related to the loss of information in the signature. A simple way to estimate the error is to count the items that are not present in the signature, i.e. the number of bits equal to 0 in the bitmap:

$$E_k(\alpha, S) = |\mathcal{I}| - \|Sig_k(\alpha, S)\| = \|\overline{Sig_k(\alpha, S)}\|$$

where $\|X\|$ represents the number of bits equal to 1 in bitmap X and $\overline{X}$ represents the complement of X. The signature of a customer's purchase sequence T is the maximal signature for a segmentation of size k:

$$Sig_k(\alpha) = Sig_k(\alpha, S_{opt}(\alpha, k)), \text{ where } S_{opt}(\alpha, k) = \underset{S \in \mathscr{S}_{n,k}}{\arg\max} \|Sig_k(\alpha, S)\|$$

The segmentation size k is given a priori, either as an integer or as a percentage of n, the size of the input sequence (similar to support count and support [1]). The latter is called the *relative number of blocks* denoted by *RNB*.

3.2 Dynamic Programming for Signature Mining by Segmentation

Now, we present an algorithm for computing signatures by sequence segmentation based on Dynamic Programming (DP). This algorithm returns an optimal solution, i.e. a maximal signature.

Bingham [10] presents a formulation of DP for sequence segmentation. It is based on a table A of size $k \times n$ where k is the size of the segmentation and n is the number of itemsets (receipts) in the input sequence α. So, rows of A represent segments and columns represent itemsets of the input sequence. Let $\alpha[j, i]$ denote the subsequence of α starting at index j and ending at index i. A cell $A[s, i]$ of table A denotes the error of segmenting the sequence $\alpha[1, i]$ using s segments, formally defined by:

$$A[s, i] = \min_{2 \leq j \leq i} (A[s-1, j-1] + E(S_{opt}(\alpha[j, i], 1))) \quad (1)$$

[5] For the sake of simplicity, we focus here on a bitmap representation. To cope with memory consumption, a more efficient representation method, such as the dynamic bit vector (DBV) architecture, could be used [13].

where $E(S_{opt}(\alpha[j,i],1))$ is the minimum error that can be obtained for the subsequence $\alpha[j,i]$ when representing it as one segment. In our case, it is simply the number of items that are present in the segment receipts. In the signature mining problem, as presented in Section 3.1, the representative of a segment is not a numeric value but a set of items. In order to compute the reconstruction error, an A cell stores the bitmaps of the best signatures obtained so far. Since several signatures may exhibit the same reconstruction error value, an A cell contains a set of bitmaps. Intuitively, $A[s,i]$ is computed by considering, for all $j \in [s,i]$, the composition of a signature obtained for an $(s-1)$-segmentation of a subsequence $\alpha[1,j-1]$, stored in $A[s-1,j-1]$, and the signature of the new segment $Sig_1(\alpha[j,i])$.

Formally, it is defined by:

$$A[s,i] = \underset{s \leq j \leq i}{\text{amaxN}}(\underset{Sig_{s-1} \in A[s-1,j-1]}{\text{amaxN}}(Sig_{s-1} \wedge Sig_1(\alpha[j,i]))) \tag{2}$$

where

$$\underset{i}{\text{amaxN}}\, P(i) \equiv \arg\max_{P(i)} |P(i)|$$

($amaxN$ returns the maximal elements of a set with respect to norm $|.|$). The representation error associated with cell $A[s,i]$ is simply $E(P)$, which is identical for every bitmap $P \in A[s,i]$. Thus, all such signatures having a maximal size are stored in $A[s,i]$.

Table 1 displays the progressive segmentation by Dynamic Programming of sequence $\alpha = \langle (ab)(abc)(acd)(abd) \rangle$. The leftmost column gives the indices of segments. The bottom row gives the indices over sequence α as well as their associated itemset and bitmap. Other cells of Table 1 details the results of operations performed by DP formalized by equations (1) and (2). m represents the error minimization operation of segmentation. Note, that in the signature mining we are looking for maximal signatures w.r.t. the number of items and thus, m is a max operator on signatures. For example, cell $[2,3]$ computes the best signature obtained by segmenting sub-sequence $\alpha[1,3]$ into 2 segments. There are 2 ways to segment $\alpha[1,3]$: $\alpha[1,2] - \alpha[3,3]$ and $\alpha[1,1] - \alpha[2,3]$. In the first case, the representative of $\alpha[3,3]$ (*i.e.* its associated bitmap) is composed with the best signature obtained for sub-sequence $\alpha[1,2]$ given by $A[1,2]$. Actually, the composition operation (denoted by $\circ$ in the table), is simply a logical AND on bitmaps. In the second case, the representative of $\alpha[2,3]$ is composed with the best signature for sub-sequence $\alpha[1,1]$ given by $A[1,1]$. The representative of several sequence elements is simply a logical OR on their associated bitmaps. The best signature for the whole sequence α and a 3-segmentation is given by $A[3,4]$.

Algorithm 1 presents a DP algorithm for sequence segmentation and signature extraction. The first row of the DP table is initialized in lines 4-6. Then rows are added iteratively until reaching the min_seg threshold. To build A_k, for $k \in [2, min\_seg[$, we just have to add the row k to A_{k-1} (lines 9-13). Finally, $A[n, min\_seg]$ provides the best signatures and related min_seg-segmentations (line 16).

Table 1. DP segmentation table for sequence $\alpha = \langle (ab)(abc)(acd)(abd) \rangle$. To be read from bottom-left to top-right.

3			m(α[3,3]$\circ$A[2,2]) =m(1011$\wedge$1100) ={1000}	m(α[4,4]$\circ$A[2,3],α[3,4]$\circ$A[2,2]) =m(1001$\wedge$1010, 1001$\wedge$1100, 1011$\wedge$1100} ={1000}
2		m(α[2,2]$\circ$A[1,1]) =m(1110$\wedge$1100) ={1100}	m(α[3,3]$\circ$A[1,2],α[2,3]$\circ$A[1,1]) =m(1011$\wedge$1110, 1111$\wedge$1100) ={1010, 1100}	m(α[4,4]$\circ$A[1,3],α[3,4]$\circ$A[1,2],α[2,4]$\circ$A[1,1]) =m(1001$\wedge$1111, 1011$\wedge$1110, 1001$\wedge$1111) ={1001, 1010, 1001}
1	m(α[1,1]) ={1100}	m(α[1,2]) ={1110}	m(α[1,3])={1111}	m(α[1,4])={1111}
	1: *(ab)* 1100	2: *(abc)* 1110	3: *(acd)* 1011	4: *(ad)* 1001

The dynamic programming algorithm has a complexity in $O(n^2k)$ and computes the optimal solution, here the maximal signature and related segmentation.

4 Experiments

In this section, we compare signatures with some other data representation models for analyzing customer purchase regularity. We demonstrate that we are able to find new regularities, that can be used to answer practical questions, such as targeted marketing.

The experiments were performed on anonymized basket data provided by a major French retailer. They were collected from may 2012 to august 2014 (27 months) from customers owning a loyalty card. To remove occasional customers, whose data do not make sense for our experiments, only customers having more than 20 baskets during the period were kept. 149 942 distinct customers, worth 16.6 GB of data, remained. The resulting database contains 3,887,979 distinct items. The retailer also provided a taxonomy that relates items to subcategories (item class). We ended up with a total of 3388 item categories. Such categories are used to get rid of minor items differences (*e.g.* packaging or brand).

4.1 Capturing purchase regularity

Mining methods that extract patterns while giving some insight of regularity, go from top-k item mining [14] to periodic pattern mining. Top-k items are the k most frequently bought items within all customer's baskets. However, top-k items do not provide an explicit information about purchase regularity. Yet, item frequency can be considered as a rough mean regularity. Periodic patterns [15] represent items that are purchased at a strict periodicity. However, some purchase delay could break the periodicity and prevent a pattern to be periodic. Signatures stand in the middle: they represent sets of items that are bought within a limited period of time and such items are bought together several times but under a non strict periodicity.

In the sequel, we compare signatures with top-k items and periodic patterns to exhibit some common and distinctive features.

Algorithm 1: Dynamic Programming for segmentation-based signature extraction

Input: $\alpha = \langle T_1, \ldots, T_n \rangle$: receipt sequence of length n, min_seg: the minimal segmentation size
Result: Sig: signatures

$A[1,1] = T_1$;
/*Initialization of the first row of A*/
for $i = 2, n$ **do**
  $A[1,i] = T_i \vee A[1,i-1]$;
end
for $s = 2, min\_seg$ **do**
  for $i = s, n$ **do**
    $Sig = \emptyset$;
    for $j = s, i$ **do**
      $Sig = Sig \cup \{A[s-1,j-1] \wedge (\bigvee_{T \in \alpha[j,i]} T)\}$;
    end
    $A[s,i] = \arg\max_{p \in Sig} |p|$;
  end
end
$Sig = A[min\_seg, n]$;
return Sig

Signatures vs top-k items In this experiment, we compare the signature content with the top-k items for each customer. We compute signatures with a relative number of blocks of 0.15 (see Section 3.1). We try different values of k for the top-k items method, and compare all of them with the signature content in Figure 1, on the left. Setting the value of k to the signature length for each customer is not possible in practive, as we do not know the signature length before hand. We therefore do not show experiments with this particualr value of k. More elaborate methods to adapt the k value to each customer, such as elbow methods [16], did not bring better results than the ones presented in Figure 1-left. The Jaccard similarity between the signature and the top-k items of most customers is between 0.5 and 0.3. This means that top-k items and signature products overlap partially. When the k value is low, the number of top-k items is significantly lower than the number of items in the signature. This leads to a low Jaccard value, even though most of the top-k items are included the signature. A similar behavior is observed for large values of k, where the top-k contains more items than the signature, leading to a low Jaccard value. For values of k close to the mean signature length: between 5 and 10 items, the Jaccard goes higher, as both sets have a similar size. Overall, the signature overlaps partially with the top-k items, and the main source of difference comes from the fact that the number of items in the signature changes for each customer, whereas it is constant for all customers in the top-k computation. The number of items in the signature could therefore be seen as a way to estimate a relevant value of k for the top-k items of a given customer. Another source of difference between

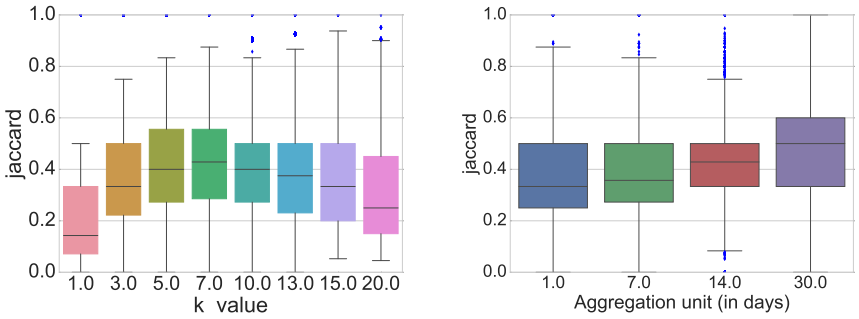

Fig. 1. On the left: Jaccard similarity between the signature and the top-k items, for different k values. This has been computed on 149 942 customers. On the right: Jaccard similarity between the signature and the longest periodic pattern, for different time scales. This was computed on 20 000 customers.

signatures and top-k items is the fact that items that are very frequently bought during a short period of time do not appear in the signature, while they are more likely to appear in the top-k items.

Periodic patterns comparison In this experiment, we compare the signature content with the periodic patterns for each customer.We used an algorithm that allows gaps between consecutive occurrences of periodic patterns [7]. As periodic patterns can only be found in a single time scale, a preprocessing step that aggregates the receipts on a given time scale (e.g., merge all receipts at the given granularity) is required. As we do not know in advance what is the relevant time scale for each customer, we are using 4 time scales to compute the periodic patterns: daily, weekly, bi-weekly and monthly purchases. For each time scale, we computed the Jaccard similarity between the longest periodic pattern and the signature. We chose the longest periodic pattern as the signature finds the longest regular pattern. If several longest periodic patterns are found, we take the one that has the largest Jaccard similarity with the signature. The results are presented in Figure 1, on the right. In this figure, we can see that the Jaccard similarity between the signature and the longest periodic pattern is mostly between 0.3 and 0.45. This means that these two sets have common elements, but still differ. Further analysis showed that the longest periodic pattern is almost totally contained in the signature. This means that the signature is composed of most items from the longest periodic pattern. This periodic part of the signature represents between one third and one half of the total signature. The remaining part of the signature contains items that are not periodic but that are regularly bought. This highlights the flexibility of the signature, as it manages to capture periodic products, while also capturing non periodic regular purchases.

Signatures capture non periodic regularities because their segments can be of arbitrary length. More specifically, each customer signature segment can contain multiple baskets, and can therefore span on different time scales. On the other

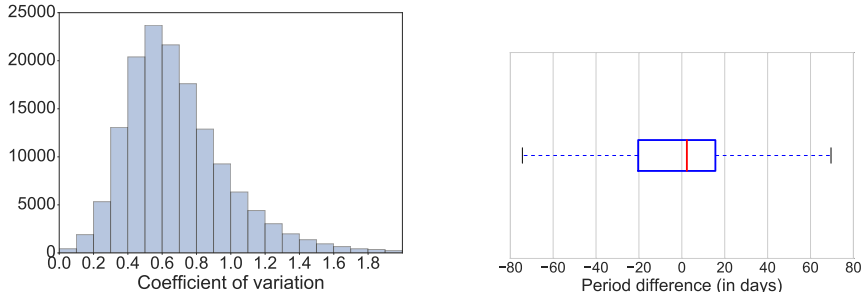

Fig. 2. On the left: Distribution of the segment length coefficient of variation of 149 942 customers. On the right: difference between the signature period and the most similar periodic pattern period.

hand, periodic patterns have a fixed segment length and cannot span on different time scales. To illustrate this difference, we plot the coefficient of variation of the segment size for each customer in Figure 2, on the left. Most customers have a coefficient of variation greater than 0.4, which means that most customers have variations in their purchase rhythms. Almost no customers have a coefficient of variation equal to zero, whereas all customers have a coefficient of variation of 0 for periodic patterns by definition. Nevertheless, the coefficient of variation remains mostly below 1, which means that customers show a regular purchase behavior. Therefore, the signature segment still captures a regular behavior of the customer. This shows that introducing flexibility in the period allows us to capture more regular products than existing methods (see Figure 1-right), while capturing a regular behavior (see Figure 2-left).

Because signatures are more flexible, their detected temporal regularity can be different than the one found by periodic patterns. To compare the period found by both methods, we compared the difference between the mean segment size of the signature, with the largest period of the periodic pattern that is the most similar with the signature (according to the Jaccard similarity). We choose the largest period of the periodic pattern, because signatures segments are as large as possible. This effect is due to the fact that segments have to cover the whole sequence. The results are presented in Figure 2, on the right. We can see that most periods found by the signature are close to the period found by the most similar periodic pattern. While there can be some differences between both periods, these differences are usually contained within a reasonable time span.

To summarize, signatures are able to find regularly purchased products, whether they are periodically bought are not. The flexibility of the regularity definition of the signatures allows us to find these products without any pre-processing step. Moreover, signatures are able to find the underlying period of customers, that is consistent with the one found by periodic patterns. Signatures therefore find the time regularity of a customer, along with the regular products. This regularity cannot be totally captured by existing methods.

4.2 Insights from Signatures

As shown in the previous section, signatures group transactions into segments to find a set of regular products, that can not be totally found by existing methods. More specifically, let us consider the case of a real customer (named A). The store visits of customer A are represented in Figure 3, on the top. For comparison, we also consider customer B whose store visits have a signature identical to her largest periodic pattern (shown in Figure 3 on the bottom). The comparison of both Figures clearly shows that the customer A has no clear buying pattern, while the customer B has a clear buying pattern: she buys her groceries almost every Saturday. Nevertheless, by computing the signature on the customer A, we are able to detect her underlying period. Indeed, her signature contains 9 products: *Biscuits, Hazelnut spread, cheese, frozen meat, pasta, cream, butter, ham* and *chocolate powder*. Only some of them are bought during the same store visit, and these purchases usually spread over 4 transactions, for a segment length of 2 weeks on average. Among these products, some of them have a periodic buying pattern (*pasta, ham* and *hazelnut spread*), while the others are bought more sporadically. Nevertheless, this whole set of products has consistency and is related to meal and break food for children. The signature was therefore able to identify the purchase rhythm (both period and products) of a customer who had no clear buying pattern when using existing methods.

Signatures can also help marketers to answer the problem of finding the most appropriate time and products to give a coupon on, for a given customer. To achieve this targeted coupon policy, it would be interesting to be able to know what kind of products this customer is likely to buy in the next visits, to be able to give this customer targeted coupons. Thanks to the signature, we can provide the marketer with information about the time and content of next purchases. Indeed, if this customer has purchased *Biscuits, cheese, frozen meat, cream* and *butter* over 2 transactions in a week, we know from the signature that this customer is likely to be buying *Hazelnut spread, pasta, ham* and *chocolate powder* in the next 2 transactions over the next week. This because we know from the signature that this customer has the habit of buying *Biscuits, Hazelnut spread, cheese, frozen meat, pasta, cream, butter, ham* and *chocolate powder* in 4 transactions over 2 weeks. As we are observing a portion of a signature segment, we can guess the products that are likely to be bought in the next week. This information is of prime interest for retailers, as they could then target their ads on the right products for each customer. It should be noted that periodic patterns would have missed the part related to break food for children, as only *pasta, ham* and *hazelnut spread* were considered periodic.

5 Conclusion

Getting a better understanding of individual customers is becoming a differentiating factor in a data-driven retail context. We have presented a novel notion of *customer signature*, that gives for each customer a good understanding of the

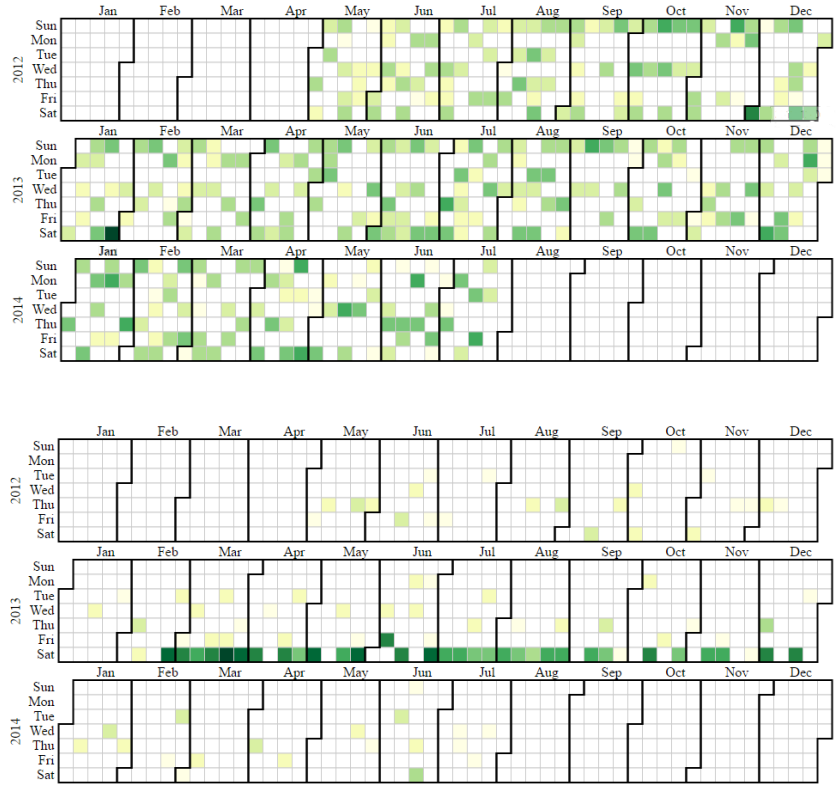

Fig. 3. Receipts of a non periodic customer (A), on the top, and periodic customer (B), on the bottom. Each green rectangle represents a visit to the store. The darker the green, the more products were bought during that visit.

products most regularly bought, as well as of the household rhythm. Our experiments have shown that this approach, thanks to its flexibility, allows to get deep insights on purchasing rhythms that are not provided by existing algorithms. The approach itself builds up on a large body of work on sequence segmentation, taking advantage of years of research on efficient exact algorithms.

This work opens new perspectives. A first one is to take product categories into account, allowing to find new types of regularities over product categories or brands. From an application point of view, with our retail partner we are investigating the use of signatures for preventive actions against churn. Another exciting perspective is to test the use of signatures on other domains than retail. Thanks to the generality of the definitions, it can be easily applied on any sequence of itemsets where a segmentation is relevant. We performed preliminary experiments on datasets of labeled TV programs, with promising results: while signatures with a high number of blocks detect regular daily programs, signatures with fewer segments but many items can detect relatively short span

events (such as Roland-Garros tennis contest) for which TV channels devote many special programs, that are picked up by the signature.

References

1. R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. 17th Int. Conf. on Management of Data*, pp. 207–216, 1993.
2. R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th Int. Conf. on Data Engineering*, pp. 3–14, 1995.
3. V. S. Tseng, B.-E. Shie, C.-W. Wu, and P. S. Yu, "Efficient algorithms for mining high utility itemsets from transactional databases," *Trans. on Knowledge and Data Engineering*, vol. 25, no. 8, pp. 1772–1786, 2013.
4. V. L. Miguéis, A. S. Camanho, and J. Falcão e Cunha, "Mining customer loyalty card programs: The improvement of service levels enabled by innovative segmentation and promotions design," in *Exploring Services Science*, pp. 83–97, 2011.
5. V. L. Miguéis, A. S. Camanho, and J. Falcão e Cunha, "Customer data mining for lifestyle segmentation," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9359–9366, 2012.
6. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *SIGMOD Int. Conf. on Management of Data*, pp. 1–12, 2000.
7. P. L. Cueva, A. Bertaux, A. Termier, J. Méhaut, and M. Santana, "Debugging embedded multimedia application traces through periodic pattern mining," in *Proc. 12th Int. Conf. on Embedded Software*, pp. 13–22, 2012.
8. H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data mining and knowledge discovery*, vol. 1, no. 3, pp. 259–289, 1997.
9. G. Casas-Garriga, "Discovering unbounded episodes in sequential data," in *Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Database*, pp. 83–94, 2003.
10. E. Bingham, "Finding segmentations of sequences," in *Inductive Databases and Constraint-Based Data Mining*, pp. 177–197, 2010.
11. E. Terzi and P. Tsaparas, "Efficient algorithms for sequence segmentation," in *Proc. SIAM Conference on Data Mining*, pp. 314–325, 2006.
12. J. Ayres, J. Flannick, J. Gehrke, and T. Yiu, "Sequential pattern mining using a bitmap representation," in *Proc. 8th Conf. on Knowledge Discovery and Data mining*, pp. 429–435, 2002.
13. B. Vo, T.-P. Hong, and B. Le, "DBV-miner: A dynamic bit-vector approach for fast mining frequent closed itemsets," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7196–7206, 2012.
14. J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining top-k frequent closed patterns without minimum support," in *Proc. Int. Conf. on Data Mining (ICDM)*, pp. 211–218, 2002.
15. S. Ma and J. L. Hellerstein, "Mining partially periodic event patterns with unknown periods," in *Proc. 17th Int. Conf. on Data Engineering*, pp. 205–214, 2001.
16. R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.